LOW-LATENCY RANDOM ACCESS TO COMPRESSED VIDEO

Background and Summary of the Invention

This invention relates to random access to a compressed video data stream (bit stream), and in particular to methodology and a system (or apparatus) for furnishing and enabling low-latency random access to such a stream.

In both the systemic and methodologic realms, the invention involves, among other things, both the preparing of a conventional video data stream for low-latency, and quickly achieved high-quality(resolution) random access, and the ultimate reception and utilization (i.e. viewing) of that prepared data stream. For the purpose of illustration herein, a preferred and best-mode embodiment of, and manner of implementing and practicing, the invention are disclosed and illustrated in the context of television channel surfing-a representative setting wherein the invention has been found to offer special utility. This context is aptly representative generally of the invention's useful applicability in the areas of video broadcasting and video streaming.

By way of general background regarding conventional understanding, due to temporal prediction, a video decoder cannot begin decoding a video data stream at a frame that is predicted from previous frames. Many applications require a user to be able enter a bit stream at any time, such as during the activity known as channel surfing between broadcast streams, and during reverse, or fast-forward, "trick" video modes. A typical technique for providing random access to a compressed bit stream involves the insertion of so-called I-frames, or intra-frames. Such I-frames are pictured and employed herein, and are also referred to as video marker frames. I-frames are typically inserted into a so-called Group Of Pictures (GOP) structure, and are coded without any prediction

from other frames. They are inserted at an interval, or rate, (R). A decoder is able to access an associated video data stream at any such inserted I-frame. The latency involved in random access in such a GOP is inversely proportional to R, while the compression performance decreases as R increases. As an example, I-frames are typically inserted periodically in MPEG-2 format to provide random access.

Another concept in the prior art involves the so-called concept of "dirty random access", regarding which an access point does not begin with a clean I-frame. Rather, in this approach, a decoder begins decoding and displaying predicted frames without having access to a complete reference frame. Blocks of the frames are intra-coded so that, after some time, a clean picture is attained.

Other approaches to dealing with random access latency have been proposed in the prior art and are generally familiar to those generally skilled in the art.

The present invention departs from prior art approaches by proposing the creation, from an "engaged" source video data stream, of a compressed and either slightly time-offset (preferable), or time-synchronized, pair of divided video data streams (called downstream-deliverable streams) that are drawn directly from the source stream, and that are characterized by possessing respective, different access latencies and resolutions. In particular, and according to a preferred manner of practicing and implementing the invention, such a source data stream is split into two data streams which are slightly time-offset with respect to one another, with one of these streams being specifically characterized with a relatively low access latency (the stream which is slightly time-delayed relative to the other stream) and a relatively low image resolution, and with the

other stream being characterized by a larger access latency, and a significantly larger, or greater, image resolution.

These two streams are preferably multiplexed and transmitted so as to be received at a receiving site where they can, in terms of certain fundamental steps practiced by a portion of the present invention, be (a) sought, (b) monitored, (c) and selected in a manner, which uniquely introduces a definitive low-latency access, followed by a rapidly achieved, high-resolution image display.

These, and various other features and advantages, which are offered and attained by the system, apparatus and methodology of the present invention will be more fully appreciated in light of the detailed description which shortly follows, when read in conjunction with the accompanying drawings.

Description of the Drawings

Fig. 1 is a block/schematic diagram illustrating a high-level "picture" of the methodology and the apparatus of a preferred and best-mode embodiment of, and manner of practicing, the present invention.

Fig. 2 illustrates schematically the characteristics of two differently characterized, downstream-deliverable video data streams which are created for multiplexed transmission in accordance with practice of the present invention.

Fig. 3 is a schematic representation of how a dual-data-stream video transmission, such as that pictured in Fig. 2, created by practice of the invention, is utilized at a receiving viewer's location by receiving apparatus which is constructed in accordance with the invention to provide definitive low-latency, quickly established high-resolution, access effectively to a source video data stream.

Figs. 4, 5 and 6 present respective different tables of information that are employed to highlight, in relation to prior-art conditions, various features of the present invention.

Detailed Description of the Invention

As mentioned above, the present invention is described herein in relation to its utility in the contexts both of video broadcasting and video streaming. The description which now follows should be read with the understanding that it is presented in a manner intended to highlight the utility of the invention in these two particular areas of video data transmission and reception. Additionally, Fig. 1 in the drawings should be viewed now as an illustration which pictures both the systemic and methodologic characteristics of the present invention.

In Fig. 1, a block 10 represents a source for supplying an uncompressed, output video data stream 12 (pictured as an arrow), which data stream has a particular, selected bandwidth which will be referred to herein as the source bandwidth. Such a data stream may emanate from any suitable video source.

Data stream 12, in accordance with practice of the present invention, is fed to a splitter (or divider, or dividing structure) 14, whose input side (its left side in Fig. 1, which is referred to herein as engaging structure) engages the source stream. With respect to such an engaged source stream, the splitter operates to split, or divide, the source data stream into two, broadcastable, derivative video data streams (downstream-deliverable data streams) which are represented by arrows 16, 18. Arrow 16, 18 are seen effectively to "bracket" a block 20 which represents suitable, conventional video-data buffer structure. These downstream-deliverable video data streams differ, in that stream 16 is a

relatively low-latency, low-resolution stream, and stream 18 is a relatively higher-latency, higher-resolution stream. As will be more fully explained shortly, each of these two data streams is made up of a series of frames which are referred to herein as I-frames, or marker frames, and P-frames, known in the art as predicted frames. The separation between I-frames in each stream is a measure of access latency, with closely spaced I-frames characterizing a data stream with relatively low latency, and more widely separated I-frames characterizing a higher-latency access characteristic.

It is in the region of what is shown in Fig. 1 which extends from source 10 to arrows 16, 18 that practice of the present invention, and apparatus which implements that practice, creates the relevant, differentiated, downstream-deliverable video data streams that enable the present invention to enhance rapid, high-resolution viewer access effectively to a source video data stream. In the region generally marked M in Fig. 1, data streams 16, 18 are multiplexed, preferably slightly time-offset relative to one another, and transmitted, in any suitable fashion, for reception at a user location, such as that generally represented at L in Fig. 1. With respect to a time offset, the low-latency stream is preferably delayed by about 1/4-second. The streams may also be synchronized if desired.

Preferably, data streams 16, 18 are appropriately compressed prior to transmission, and any appropriate, conventional compression technique, or techniques, may be employed for this purpose. Preferably also, the combined bandwidth resource called for by data streams 16, 18 for transmission is about the same as that which would be required to transmit a more conventional, single video data stream which is characterized by relatively rapid access and relatively high resolution. Obviously, the

particular selected resolutions and I-frame placements (spacing) chosen for data streams 16, 18 will determine this combined data-stream bandwidth requirement. These are matters of user choice, and are not specifically critical to practice of the invention

While, as the case is here, the specific approaches that are employed regarding source data stream 12 to effect splitting and compression, and to assure establishment of the preferred bandwidth-utilization characteristic just mentioned, may be entirely conventional, and thus are not discussed in any further or greater detail herein, the act of splitting *per se* to create the two, mentioned, latency and resolution-differentiated, downstream-deliverable video data streams is unique, and forms an important part and contribution of the present invention. This preparation from a source data stream preferably takes place, as has been generally expressed with respect to the description given so far for Fig. 1, at, essentially, the location from which video data is to be broadcast to viewers.

Continuing with Fig. 1, downstream from where data-stream preparation, as just outlined, takes place by operation of the present invention, and preferably at the site of a viewer's television receiver, for example at location L in Fig. 1, a searching (seeking) function, represented in Fig.1 by a block 22, takes place. Block 22 is thus referred to herein as seeking structure. This searching/seeking function, which occurs in accordance with practice of the invention, is initiated, for example, by a viewer's undertaking a random-access channel-surfing activity, which activity is represented by a block 24 labeled "Start" in Fig. 1. Specifically what takes place in block 22, with respect to the start of a searching or seeking function (in accordance with the invention), will be more fully explained shortly, but for now it is sufficient simply to state that this function

initiates quick (very low latency) presentation of uncompressed video data on the screen of the viewer's television receiver. As will also be more fully explained shortly, the searching/seeking function initiated by block 24, and performed by block 22, also implements certain monitoring and selection functions which are undertaken by block 22 in cooperation with a downstream switching block 26 shown in Fig. 1. Block 22 is also referred to herein as monitoring structure, and block 26 as selecting structure.

Disposed intermediate blocks 22, 26 in Fig. 1 is a block 28 which implements appropriate, conventional data-buffering activity between blocks 22, 26.

Operations of the initiation, seeking, monitoring, selecting and switching functions results in a furnishing to the user's television receiver at site L of an appropriate uncompressed output video presentation (or output video data stream) which, in Fig. 1, is represented by broad arrow 30.

Turning attention now to Fig. 2 in the drawings, this figure schematically illustrates data streams 16 and 18. More specifically, Fig. 2 provides an illustration of representative segments, or lengths, of these two data streams, with certain symbology and graphical techniques employed to highlight the differences between these two video data streams. Thus, data stream 16 which, as mentioned earlier, is constructed to be characterized by low latency and relatively low resolution, is pictured as an alternating series of P-frames and I-frames, such as the four P-frames which are shown at 16a, 16b, 16c, 16d, which P-frames alternate, on a one-to-one illustrative basis, with I-frames, such as those shown as 16e, 16f., 16g, 16h. It is thus seen that the I-frames in data stream 16 are spaced at extremely close intervals, and it is this placement of these frames which causes data stream 16 to be characterized with very low access latency.

The individual I-frames and P-frames in data stream 16 are represented in Fig. 2 as small shaded parallelograms, and the size of these parallelograms, in relation to the size of similar parallelograms appearing in data stream 18 (still to be discussed), is intended to reflect the fact that data stream 16 is characterized with a relatively low resolution.

Data stream 18, by way of contrast, includes both I-frames and P-frames organized in a fashion wherein a significant number (which is not specifically illustrated in Fig. 2) of P-frames resides between each two next-adjacent I-frames, only one of which is shown in Fig. 2. Thus illustrated for data stream 18 in Fig. 2 are a string of P-frames 18a, 18b, 18c, 18d, 18e, 18f, 18g which are disposed to the right of the single illustrated I-frame 18h. Trailing to the left side of frame 18h in Fig. 2 are three more P-frames 18i, 18j, 18k. As was mentioned just above, the various I and P-frames in data stream 18 are represented by shaded parallelograms which are larger than those which represent the frames in streams 16. These larger rectangles are used to indicate that the resolution which characterizes data stream 18 is higher than that which characterizes data stream 16. The access latency which characterized data stream 18 is significantly higher also because of the fact that a significantly larger (unknown number) of P-frames resides between each next-adjacent pair of I-frames.

In Fig. 2, the overall left-to-right lengths of the two illustrated data streams are displayed in a fashion to indicate how they relate in time relative to one another over the same general time interval, which is also measured in a left-to-right manner in Fig. 2. As has already been mentioned, stream 16 is preferably delayed (time-offset) relative to stream 18 by an interval of about 1/4-second. If desired, the streams may also be

synchronized. Fig. 2 should be viewed as illustrating, generally, both of these approaches. At the right side in Fig. 2 is a dash-dot line labeled TD which is intended to represent what is referred to herein as a time datum with respect to which, during transmission, one can imagine that the individual frames in the two data streams pass as time progresses. In other words, with respect to the frames that are contained in data stream 16, the first frame to pass this time datum is frame 16a, the next 16e, the next 16b, and so on. Similarly, the first frame in data stream 18 to pass the time datum line is frame 18a, followed by frame 18b, followed by frame 18c, and so on.

Fig. 2 thus provides a representation of the way in which the two, multiplexed, divided data streams produced in accordance with practice of this invention flow as downstream-deliverable data streams 16, 18 toward a viewer's site, such as site L in Fig. 1.

According to practice of the invention, when a user seeks access to the image information contained in data streams 16, 18, he or she does this by implementing the seek or start function represented by block 24, whereupon block 22 begins to monitor and examine received data streams 16, 18 for the purpose of detecting the very first I-frame in either of the two data streams which effectively passes a time datum mark, such as dash-dot line TD in Fig. 2. As can be seen in Fig. 2, and with respect to the portions of data streams 16, 18 which are pictured in that figure, the very first I-frame which will be encountered will be frame 16e in low-latency data stream 16.

On this detection of an I-frame occurring, block 26 effectively directs into output signal 30, for presentation on the viewer's reception screen, the low-latency, low-resolution imagery data represented by, and contained within, data stream 16. Inasmuch

as the first I-frame encountered has occurred in the low-latency data stream, the searching, monitoring and switching functions implementable by blocks 22, 26 remain active, with block 22 continuing now to search for the next-occurring I-frame which appears in the higher-latency, higher-resolution data stream 18. This "next-occurring" frame, which will be, in accordance with what is pictured in Fig. 2, frame 18h, will "appear" after frames 16a, 16e, 16b, 16f, 16c, 16g, 16d have passed the time datum line. When block 22 then detects the arrival of I-frame 18h in the higher-latency, higher-resolution data stream, it effectively invokes a switching function in block 26 which causes the output signal represented by arrow 30 now to switch to presenting the full content of higher-resolution data stream 18. At this point in time, the searching monitoring and switching functions are concluded, and the viewer is presented with full resolution imagery.

Fig. 3 provides a graphical representation of the content of output signal 30 which exists as a consequence of the searching and switching operations which have just been described. Here, one sees that the order of frame presentation to the screen at site L is 16a, 16e, 16b, 16f, 16c, 16g, 16d, 18h, 18i, 17j, 18k. This sequence of frame presentation clearly demonstrates the powerful low-latency access to video data offered by practice of this invention, followed rapidly by full-resolution image presentation.

Had the first-encountered I-frame been such a frame in data stream 18, output signal 30 would have immediately been derived from this higher-resolution stream, and the activities involving seeking, monitoring and selecting would have been immediately terminated.

Thus, and reviewing now, as a follow-up to the discussion above, various considerations relating (in the context of conventional practice) to the implementation of this invention, a video bit stream composed of intra frames (I-frames) and predicted frames such as P-frames can be accessed at an I-frame only. To control access latency, I-frames are periodically inserted. The I-frame period determines the access latency. If T is the time between I-frames, the access latency corresponds to a random variable uniformly distributed on the interval [O,T). The statistical values which describe this situation are presented in the table shown in Fig. 4.

An I-frame of 1-second gives a maximum access latency of 1-second, and an average access latency of 500-ms. An I-frame period of 0.1-seconds reduces these numbers by a factor of ten. The compromise is bitrate. Decreasing the I-frame period increases the required bitrate for the same output visual quality. In practice an I-frame period is preferably chosen to be near 1-second.

Regarding the structures of the herein proposed two, downstream-deliverable video data streams, such two video streams are used effectively to reduce access latency. Four parameters are used in the encoding of each stream -- namely (a) resolution, (b) frame rate, (c) signal-to-noise ratio (bitrate), and (d) access latency. The low-latency stream is coded in a fashion which compromises resolution, frame rate, and signal-to-noise ratio in exchange for greatly improved (shortened) access latency. The parameters selected for the higher-resolution stream are chosen such that that stream is "accessed", delivered video-signal quality is excellent. Sample values of appropriate "dual-stream" parameters, and of related, representative access latencies, are shown in the tables presented in Fig. 5 and 6.

The invention thus proposes an efficient, effective, and relatively simple method and apparatus for improving, by minimizing, access latency to a high-resolution video data stream. Utilizing the approach of splitting a source data stream into two latency- and resolution-differentiated data streams for transmission, the method and apparatus of the invention offer low-latency access time wherein the maximum delay (latency) at a viewer's site is never greater than the time distance between I-frames in the low-latency, low-resolution stream. A high-resolution image is presented to a viewer, in all cases, just as soon (after the viewer requests access) as the higher-latency, higher-resolution stream next presents an I-frame. Transmission of the two, proposed, divided video data streams can be accomplished without taxing available bandwidth resources, and specifically by constructing these two data streams in such a manner that, collectively, they require only about the same bandwidth as that required by a typical high-resolution, modest access-latency, single data stream.

Those generally skilled in the art will understand that, while a preferred and best-mode embodiment of the invention has been described and illustrated herein, and a modification mentioned regarding transmission of the proposed, two, divided streams in a synchronized manner, other variations and modifications are possible that come within the scope of the invention.